IL99/00057

PCT/IL 99/00057
18 MAY 1999

09/601258

REC'D 28 MAY 1999
WIPO PCT

# THE UNITED STATES OF AMERICA

**TO ALL TO WHOM THESE PRESENTS SHALL COME:**

**UNITED STATES DEPARTMENT OF COMMERCE**
**United States Patent and Trademark Office**

**February 5, 1999**

**THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE UNDER 35 USC 111.**

**APPLICATION NUMBER:** ***60/072,977***
**FILING DATE:** ***January 29, 1998***

**PRIORITY DOCUMENT**
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

**By Authority of the**
**COMMISSIONER OF PATENTS AND TRADEMARKS**

**MARGARET BASSFORD**
**Certifying Officer**

FOLEY & LARDNER
3000 K Street, N.W., Suite 500
P.O. Box 25696
Washington, D.C. 20007-8696
(202) 672-5300

# PROVISIONAL APPLICATION FOR PATENT

Assistant Commissioner for Patents
Washington, D. C. 20231

Sir:

This is a request for filing a **PROVISIONAL APPLICATION FOR PATENT** UNDER 37 CFR 1.53(c).

| INVENTOR(S)/APPLICANT(S) | | | |
|---|---|---|---|
| LAST NAME | FIRST NAME | MIDDLE INITIAL | RESIDENCE (City & either State or Country) |
| LINIAL | Michal | | Jerusalem, Israel |
| LINIAL | Nathan | | Jerusalem, Israel |
| TISHBY | Naftali | | Jerusalem, Isreal |
| YONA | Golan | | Jerusalem, Isreal |
| | | | |

| TITLE OF THE INVENTION |
|---|
| AN AUTOMATIC GRADED PARTITIONING OF ALL KNOWN PROTEINS |

In connection with this application, the following are enclosed:

11 Pages of Specification (Optional: ☒ Abstract ☐ Claims _)

7 Sheets of Drawings

__ Assignment to:

__ Statement of Small Entity Status

XX Other: Check for $150.00

Attorney Docket No. 023826/0109

The fee has been calculated as shown below. (Small entity fees indicated in parentheses.)

| | | |
|---|---|---|
| Filing Fee | $150 ($75) | $150.00 |
| Rule 17(k) fee for non-English text | $130 | 0.00 |
| Assignment Recording Fee | $ 40 | 0.00 |
| | TOTAL FEE | $150.00 |

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.

☒ No ☐ Yes, the name of the U.S. Government agency and the Government contract number are: .

A check in the amount of the above TOTAL FEE is attached. The Commissioner is hereby authorized to charge any deficiency or credit any overpayment to Deposit Account No. 19-0741.

Respectfully submitted,

Date: January 29, 1998
Docket No.: 023826/0109

Bernhard D. Saxe
Reg. No. 28,665

# AN AUTOMATIC GRADED PARTITIONING OF ALL KNOWN PROTEINS

Michal Linial[1], Nathan Linial[2], Naftali Tishby[2] and Golan Yona[1*]

[1]Department of Biological Chemistry Institute of Life Sciences. Hebrew University. Jerusalem 91904, Israel.

[2]Institute of Computer Science. Hebrew University, Jerusalem 91904, Israel.

*Corresponding author. Tel +972-2-6585775. Fax +972-2-6585439. email: golany@cs.huji.ac.il

Abstract

We investigate the space of all protein sequences. We combine the standard measures of similarity, to associate with each sequence an exhaustive list of neighboring sequences. These lists induce a (weighted directed) graph whose vertices are the sequences. The weight of an edge connecting two sequences represents their degree of similarity. This graph encodes much of the fundamental properties of the sequence space. The idea that underlies our work is that interesting homologies among proteins can be deduced by transitivity.

If we eliminate all edges of weight below a certain significance threshold, the graph splits into connected components. These automatically induced sets of proteins are closely correlated with natural biological families. By performing this procedure at varying thresholds, in a stepwise manner, we obtain a hierarchical organization of the connected components, and thus of all known proteins.

The results show that this method successfully identifies many biological families. By varying the threshold of statistical significance, we discover finer sub-families that make up known families of proteins. Likewise, this procedure exposes linkages between distinct protein families. Broadly speaking, protein families turn out to be connected in two distinct ways: (i) Through multi-domain proteins, each of which is associated with a distinct protein family, or (ii) Through proteins much of whose sequence is shared by the two families. The latter may be considered as linkers or ancestor proteins. Consequently, many interesting relations between protein families are revealed and hierarchical organization within protein families suggest themselves.

An interactive web site including the results of our analysis has been constructed, and is now accessible through http://protomap.cs.huji.ac.il

## 1 Introduction

In recent years we have been witnessing a constant flow of new biological data. Large-scale sequencing projects throughout the world turn out new sequences, and create new challenges for researches. Many sequences that are added to the databases are unannotated and await analysis. Currently, 12 complete genomes (of yeast, E. Coli, and other bacteria) are available. About 35%-50% of their proteins have an unknown function [1].

In the absence of structural data, the analysis necessarily starts by investigating the sequence proper. Sequence analysis has many aspects: composition, hydrophobicity, charge, secondary structure propensity and more. The most effective analyses compare the sequence under study with the whole database, in search for close relatives. The properties of a new protein sequence are extrapolated from those of its neighbors. Since the early 70's, algorithms were developed for the purpose of comparing protein sequences efficiently and reliably [2]-[6].

It is generally claimed that two sequences with over 30% identity along much of the sequences, are very likely to have the same fold [7]-[9]. Proteins of the same fold usually have similar biological functions. Nevertheless, one encounters many cases of high similarity in fold, despite a low sequence similarity [10]. Such instances are, unfortunately, missed by simple searches against the database.

Significant sequence similarity entails, with a high statistically confidence, a homology among the proteins in question. By definition, homologous proteins evolved from the same ancestor protein. The degree of conservation varies among protein families. However, homologous proteins almost always have the same fold [11]. Homology is clearly a transitive relation: If A is homologous to B, and B is homologous to C, then A is homologous to C. This simple observation can be very effective in discovering homology. This is particularly useful in the so called "twilight zone" [12], where sequences are identical to, say, 10-25%. Transitivity can be used to detect related proteins, beyond the power of a direct search.

This observation has been made before. In [13] transitivity and single linkage clustering are employed to extract similarities among 2000 E. Coli protein sequences. In [14] a similar analysis is performed on 75% of the proteins encoded in the E. Coli genome. The power of transitivity in inferring homology among distantly related proteins (e.g., Streptomyces griseus protease A and protease B) is demonstrated in [10]. In [15] transitivity was combined to a search through the database, for the purpose of modeling a protein family, starting from a single sequence. However, the full power of this idea still has not yet been exploited.

The transitive closure of the similarity relation among proteins, splits the space of all protein sequences into connected components or clusters. These are proper subsets of the whole database wherein every two members are either directly or transitively related. These sets are maximal in this respect and cannot be expanded. Thus they offer a self-organized classification of all protein sequences in the database.

The work we describe here concerns these clusters, as well as their correspondence with high level features of proteins (family, function and structure characterization). This approach leads to the definition of a new metric on the space of all protein sequences. We believe that this emerging metric is more sensitive than existing measures. Such metrics are necessary in the quest of a global self organization of all protein sequences, as discussed in [16].

## 2 Methods

This section contains a description of our computational procedure. The procedure was performed for the swissprot database [17] release 33, with total of 52205 proteins.

### 2.1 Defining the graph

We represent the space of protein sequences by means of a directed graph. The vertices of this graph are the protein sequences. Edges between the vertices are weighted with weights that reflect the distance or dissimilarity between the corresponding sequences, i.e. high similarity translates to a small weight (or distance). To compute the weight of the directed edge from $A$ to $B$, one compares $A$ against all sequences in the swissprot database, and obtains the distribution of its score. The weight is taken as the expectation value [18] of the similarity score between $A$ and $B$, based on this distribution. This is a statistical estimate for the number of occurrences of the appropriate score at a random setup, assuming the existing amino acid composition[1]. When the similarity score is statistically insignificant, the corresponding edge is discarded (details below). In other words, an edge among sequence $A$ and $B$ indicates that the corresponding proteins are likely to be related.

This graph has been constructed, using all currently known measures of similarity between protein sequences: Smith Waterman (SW) [4], FASTA [5] and BLAST [6]. These methods are in daily use by biologists, for comparing sequences against the databases. Though SW tends to give the best results on average, it is not uncommon that FASTA or BLAST are more informative [19]. Therefore we chose to incorporate all three methods into our graph, to achieve maximum sensitivity [2].

Searches may be strongly biased when the amino acid composition of the query sequence differs markedly from the overall average composition. A case in point are the effects of low complexity segments within the sequence [18]. Therefore, we also consulted the results of BLAST following a filtering of the query sequence, to exclude low complexity segments (using the SEG program [21]).

The following sections contain a detailed description of the procedure of assigning weights to edges. The procedure starts from creating the neighbors list for each sequence, in each of the three methods. A numerical normalization is applied first to all methods, so they are all on comparable scales. Then, only statistical significant similarities are maintained in these lists. Finally, the weight of an edge is defined as the minimum associated to it by any of the three methods, to capture the apparently strongest relation.

[1] A high expectation value corresponds to a weak connection.

[2] In order to better identify remote homologous, we used FASTA with the BLOSUM50 scoring matrix [20], while SW and BLAST used the BLOSUM62 scoring matrix.

## 2.2 Scaling all methods to a single scale

It is relatively easy to compare between scores that a particular method assigns to different comparisons. However, how does one compare between scores that are assigned by different methods? We performed the following calculation: Pick any protein, carry out an exhaustive comparison against the whole database and consider the highest scores in each of the methods. Now plot these values and compare two methods at a time. These scores show a very strong linear relation in log-log scale (not shown), therefore introducing a (usually small) multiplicative factor, per each protein and per method, scale the three methods to a single reference line[3].

## 2.3 Defining the neighbors' list

It is, of course, very difficult to set a clear dividing line between true homologies and chance similarities. Expectation values below $10^{-3}$ can be safely considered significant and those above 10 reflect almost pure chance similarity. However, the range within is difficult to characterize, and truly related proteins may have expectation values around 1. An overly strict threshold will miss important similarities within the twilight zone, whereas an excessively liberal criterion will create many false connections. The exact threshold for each method was set to best discern among related and unrelated proteins. Our choice is based on the overall distribution of distances over the entire protein space, as given by each of the three methods.

This is illustrated in Fig. 1, which shows the distribution of expectation values over the entire swissprot database, for SW, FASTA, and BLAST. The graphs in Fig. 1 naturally suggest a threshold for each method. The distribution drawn a log-log scale is nearly linear, at low expectation values, but starts a rapid increase at a certain value. This value is set to be the threshold. The thresholds for SW, FASTA and BLAST are set at 0.1, 0.1 and $10^{-3}$ respectively[4]. An edge from vertex $A$ to vertex $B$ is maintained only if a significant score is obtained on comparing the corresponding proteins. Namely, if either SW or FASTA yield an expectation value $\leq 0.1$ or BLAST's expectation value is $\leq 10^{-3}$.

[3] The differences between FASTA and SW are mostly due to the different scoring matrices that are being used, and can be corrected by multiplying the original score by the relative entropy of the two matrices [22]. The differences between SW and BLAST may be due to approximations in estimating the parameters $\lambda$ and $K$ [23]. The underlying assumption in calculating these parameters is that the amino acid composition of the query sequence is close to the overall distribution. This assumption often fails, e.g. for low complexity segments. Moreover, these parameters are based on first order statistics of the sequence, the scoring matrix and the database. The corrections that are required to match SW and BLAST may be due to inaccurate approximations of the estimated parameters, or by higher order statistics of the sequence.

[4] However, if filtering leads to a significant reduction in the number of high scoring hits, a more stringent threshold is set for BLAST at $10^{-6}$.

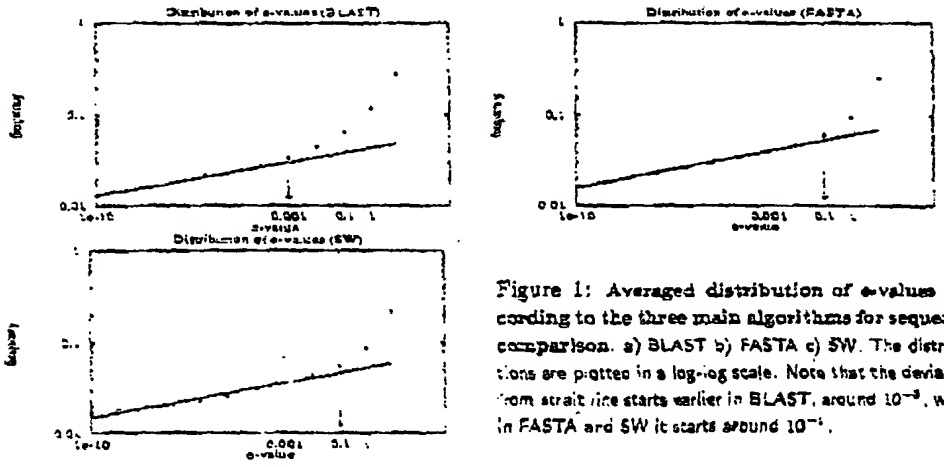



Figure 1: Averaged distribution of e-values according to the three main algorithms for sequence comparison. a) BLAST b) FASTA c) SW. The distributions are plotted in a log-log scale. Note that the deviation from strait line starts earlier in BLAST, around $10^{-3}$, while in FASTA and SW it starts around $10^{-1}$.

A major difference between BLAST and SW/FASTA is that BLAST charges no gap penalties. Consequently, BLAST tends to overestimate the statistical significance of alignments. We counter this behavior of BLAST by the above asymmetry in selecting the edges. While this property may help BLAST reveal significant similarities that the other methods miss (e.g. [19]), we have to beware highly fragmentary alignments that cannot be considered biologically meaningful. Therefore, we ignore those BLAST scores that come from a large number of HSPs (high scoring pairs), whereas the MSP (maximal segment pair) is insignificant[5].

Finally, even if the comparisons between proteins $A$ and $B$ fail to satisfy the previous criteria, the edge from $A$ to $B$ is maintained when all three methods yield an expectation value $\leq 1$.

## 3 Exploring the connectivity

We next turn our attention to the connected components of the graph we created. The transitive closure of the similarity relation among proteins, splits the space of all protein sequences into connected components or clusters. These are proper subsets of the whole database wherein every two members are either directly or transitively related. These sets are maximal in this respect and cannot be expanded. Thus they offer a self-organized classification of all protein sequences in the database. These connected components can be expected to correlate with known biological

[5] Specifically, the average and the standard deviation of the number of HSPs and the score of the MSP are calculated for high scoring sequences ($\mu_{HSP}$, $\sigma_{HSP}$, $\mu_{MSP}$ and $\sigma_{MSP}$ respectively). Those hits that are based on number of HSPs $> \mu_{HSP} + \sigma_{HSP}$, with MSP score $< \mu_{MSP} - \sigma_{MSP}$, and are not significant according to SW and FASTA are ignored.

families. The method and the results suggest that these connected components are indeed very informative and robust.

Note that this is a *directed* graph, so it is not necessarily symmetric. Specifically, it may (and does) happen that there is an edge from protein A to protein B, but none in the reverse direction. Furthermore, even if both edges exist, their weights may differ. Therefore, our notion of a component is that of a *strong connected component*[6]. The partition into strongly connected components is thus more refined that the partition into connected components.

This analysis can be performed at different thresholds, or confidence levels, to obtain an hierarchical organization. Several connected components of a given threshold may fuse together at a more permissive threshold. The analysis starts at the $10^{-100}$ threshold. Subsequent runs are carried out for $10^{-95}, 10^{-90}, \ldots 10^{-0} = 1$.

## 4 Results

Almost all of the clusters we found are meaningful. Some correspond to well known families, but many others correspond to less studied families. There are clusters that consist exclusively of unknown proteins or hypothetical proteins. An interactive web site including the results of our analysis has been constructed (http://protomap.cs.huji.ac.il). This site will help the user get acquainted with this new map of the protein space.

Table 1 shows the distribution of cluster sizes at the different confidence levels. At each level, the universe of all proteins splits into connected components (clusters). These clusters become larger and coarser with the decrease of confidence levels. Consequently, the number of isolated proteins (clusters of size 1) decrease. Note the sharp decline in the number of midsize clusters, as confidence level decrease to $10^{-0}$. Chance similarities tend to blur the picture, and cause an "avalanche", where (possibly unrelated) many families are joined to few giant clusters.

In what follows we focus only on a small number of examples. The examples are based on the observation that connected components of a given threshold may fuse together at a more permissive threshold. This fusion reflects the existence of sub-families within a family, or families within a super-family.

### 4.1 Hierarchical organization within protein families

In the next two examples we propose hierarchical organization within known families. This organization is based on the information extracted while moving across the different levels of the tree

[6] A directed graph is *strongly connected* if for every two vertices there is a directed path from x to y as well as from y to x.

| Level | over 100 | 61-100 | 21-60 | 11-20 | 6-10 | 2-5 | 1 | total no |
|---|---|---|---|---|---|---|---|---|
| $10^{-30}$ | 8 | 18 | 80 | 234 | 528 | 3727 | 29870 | 34475 |
| $10^{-35}$ | 8 | 19 | 101 | 228 | 536 | 3806 | 29078 | 33786 |
| $10^{-40}$ | 8 | 20 | 112 | 234 | 544 | 3866 | 28212 | 33018 |
| $10^{-45}$ | 8 | 23 | 119 | 261 | 565 | 3997 | 27178 | 32151 |
| $10^{-50}$ | 8 | 25 | 136 | 258 | 594 | 4060 | 26127 | 31208 |
| $10^{-55}$ | 9 | 33 | 134 | 268 | 619 | 4113 | 25030 | 30206 |
| $10^{-60}$ | 11 | 35 | 140 | 285 | 645 | 4117 | 23912 | 29145 |
| $10^{-65}$ | 12 | 35 | 136 | 321 | 650 | 4154 | 22872 | 28180 |
| $10^{-70}$ | 15 | 36 | 166 | 311 | 671 | 4134 | 21721 | 27054 |
| $10^{-75}$ | 16 | 42 | 176 | 321 | 670 | 4143 | 20584 | 25852 |
| $10^{-80}$ | 17 | 50 | 180 | 349 | 669 | 4140 | 19386 | 24785 |
| $10^{-85}$ | 21 | 52 | 190 | 359 | 673 | 4108 | 18174 | 23578 |
| $10^{-90}$ | 26 | 58 | 192 | 358 | 690 | 4034 | 17009 | 22375 |
| $10^{-95}$ | 30 | 55 | 199 | 374 | 718 | 4012 | 15656 | 21044 |
| $10^{-100}$ | 32 | 50 | 208 | 382 | 736 | 3903 | 14268 | 19582 |
| $10^{-105}$ | 36 | 54 | 224 | 378 | 719 | 3725 | 12962 | 18096 |
| $10^{-110}$ | 36 | 60 | 237 | 375 | 701 | 3538 | 11593 | 16538 |
| $10^{-115}$ | 37 | 57 | 230 | 378 | 685 | 3222 | 10221 | 14830 |
| $10^{-120}$ | 35 | 52 | 227 | 349 | 655 | 2855 | 8708 | 12882 |
| $10^{-5}$ | 24 | 41 | 191 | 282 | 528 | 2382 | 6545 | 10263 |
| $10^{-0}$ | 1 | 0 | 33 | 68 | 207 | 1292 | 4639 | 6260 |

Table 1: Distribution of clusters by their size at each confidence level.

(Scanning the hierarchy over all levels).

### 4.1.1 The small G-protein/Ras super family

The ras gene is one of a family of genes, that have been found in tumor virus genomes, and are responsible for the ability of the viruses to cause tumors in the cells they infect. In most cases this viral oncogene is closely related to a cellular counterpart (called proto-oncogene). Infection by a retrovirus carrying a mutant form of the ras gene (ras oncogene), or mutations, can cause cell transformation. Indeed, mutations in ras gene are linked to many human cancer.

The cellular ras protein bind guanine nucleotide and poesess a GTPase activity. It participates in the regulation of cellular metabolism, survival and differentiation. In the last decade many additional proteins related to ras were discovered. They all share the guanine nucleotide binding site and are of 21-30 KDa in length. They referred to as small-G-protein super-family [25].

This family of proteins composed of few sub-families: ras, rab, ran, rho, ral, and smaller sub-families. Similar to ras, these proteins participate in cell regulation process, such as vesicle trafficking (rab) and cytoskeleton organization (rho). In figure Fig. 2 we depict the relations within this family, based on the hierarchical organization obtained by our analysis. Total of 866 proteins, all belong to the small G-protein super-family, are presented. Small clusters, which correspond

to subfamilies, are formed at the high levels of confidences, and fuse to larger clusters, when the threshold is lowered. At the low level of confidence of $10^{-10}$, this family unite with the ADP-ribosylation factors family and guanine nucleotide-binding proteins, all of which are GTP-binding proteins. to form one cluster. The homology with other G proteins can be traced after screening the chance similarities at the level of $10^{-9}$ (work in progress).

#### 4.1.2 The ATP-binding transporters family

Transporters are membranous elements which provide the mechanism by which components cross the lipid bilayer within cell compartments and from its environment. The large variety of components, environmental conditions and organisms make this super-family very complex and rich [26]. This diverged family is another example for which an hierarchical organization is proposed (Fig. 3).

The sub-classification distinguishes between amino-acids transporters, oligopeptide transporters, metal transporters, multidrug resistance proteins, and many more subgroups. Out of 296 proteins presented here. 75 are hypothetical transporters which can be classified based on this organization, according to their position in the tree.

### 4.2 Relations between protein families

In the next three examples we demonstrate how the transitivity can be used to verify the relation between different, but functionally related, protein families. In these examples we focus on the connections created when moving from one confidence level to the next level.

#### 4.2.1 Super-family of motor proteins

In some cases the connection between functionally related proteins is revealed only through the connection with hypothetical proteins. One such example is the connection between the myosins and the kinesins (Fig. 4). The isolated sets (at the level of $10^{-35}$) of kinesin and kinesin-like proteins (sets C1,D1,D2), myosins (set A1), axoneme-associated proteins (set C2), and trichohyalin (set B2), were grouped together, in some cases via connections with hypothetical proteins (sets B1,C3), to form a super family of motor proteins (at the level of $10^{-30}$) with total of 120 proteins. All proteins share elongated structure, and energy dependent motor activity and are expressed throughout the evolutionary tree. They do vary in their directionality of action, tissue specificity and their highly diverged biological contexts.

#### 4.2.2 Proteins involved in the biosynthesis of complex sugars

Chitin synthase and cellulose synthase which play a major role in cell wall biogenesis, nodulation protein which are involved in the synthesis of a tetrasaccharide, and succinoglycan biosynthesis proteins which are involved in the exopolysaccharide biosynthesis, all share biological activity, in complex sugars biosynthesis.

Indeed, a relation between those families is established in our organization by lowering the confidance level from $10^{-10}$ to $10^{-5}$ (Fig. 5). As in the previous examples, the connection is established via hypothetical proteins, based on weak alignments (Fig. 6). However, the basic biological feature which characterize all these proteins, makes the connections inevitable.

#### 4.2.3 Methylases and methyltransferases

This family is another example for the importance of hypothetical proteins as linkage proteins. Through such links a natural connection between related biological families is established, as demonstrated for methylases and methyltransferases (Fig. 7).

Total of 80 proteins in 28 isolated sets (at the level of $10^{-10}$) were connected to one cluster at the level of $10^{-3}$. The Y-axes of the graph represents 11 orders of transitivity (labeled A-K). At the bottom end (A1,B1,B2,C1) as well as at the top end (J2) methylases and methyltransferases are common. Many hypothetical proteins are scattered within these clusters.

Some clusters contains exclusively hypothetical proteins (E2,H1,I2). The connection between the two ends of this graph is made through such clusters (H1,I2). The connections are based on very sparse pairwise alignments, and raise a reasonable doubt on the biological significance (Fig. 8). Yet, proteins at the two ends of the graph exhibit close biological function, therefore verify the validity of these connections.

## 5 Discussion

In this paper we address the problem of identifying high order features within the sequence space. We begin by representing the sequence space as a weighted directed graph. We explore the properties of this graph to obtain a better understanding of the space of all protein sequences. We [illegible] constituents of strong connected components to explore the constituents of this space, and their correspondence with known biological families. [illegible] at different thresholds, to obtain a hierarchical organization. This organization, reveals interesting relations between and within protein families.

Two kinds of connections between families become apparent (i) Multi-domain proteins, each domain of which is associated with a different protein family, or (ii) Connections through proteins

much of whose sequence is similar to two distinct families. The latter may be considered linkers or ancestor proteins (examples of the two types will be described elsewhere).

An interactive web site including the results of our analysis has been constructed (http://protomap.cs.h ; At this version we chose not to eliminate any potential connections. It would be interesting to receive users' feedback on which are real connections and which artifact that ought to be discarded.

## 6 Acknowledgments

We thank Hanah Margalit for many valuable discussions.

## References

[1] Pennisi, E. (1997). Microbial genomes come tumbling in. *Science* 277, 1433.

[2] Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443-453.

[3] Sellers, P. H. (1974). On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.* 26, 787-793

[4] Smith, T. F. & Waterman, M. S. (1981). Comparison of Biosequences. *Adv. in Appl. Math.* 2, 482-489 .

[5] Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity. *Science* 227, 1435-1441.

[6] Altschul, S. F., Carrol, R. J. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410.

[7] Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9, 56-68.

[8] Flores, T. P., Orengo, C. A., Moss, D. & Thornton, J. M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Protein Science* 2, 1811-1826.

[9] Hilbert, M., Bohm, G. & Jaenicke, R. (1993). Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins* 17, 138-151.

[10] Pearson, W. R. (1997). Identifying distantly related protein sequences. *Cabios* 13:4, 325-332.

[11] Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol* 266, pp 227-258.

[12] Doolittle, R. F. (1992). Reconstructing history with amino acid sequences. *Protein Science* 1, 191-200.

[13] Watanabe, H. & Otsuka, J. (1995). A comprehensive representation of extensive similarity linkage between large numbers of proteins. *Cabios* 11:2, 153-166.

[14] Koonin, E. V., Tatusov, R. L. & Rudd, K. E. (1996). Protein sequence comparison at genome scale. *methods Enzymol* 266, 295-321.

[15] Neuwald, A. F., Liu, J. S., Lipman, D. J.& Lawrence, C. E. (1997). Extracting protein alignment models from the sequence database. *NAR* **25**:8, 1665-1677.

[16] Linial, M., Linial, N., Tishby, N. & Yona, G. (1997). Global self organization of all known protein sequences reveals inherent biological signatures. *JMB* **268**, 539-556.

[17] Bairoch, A. & Boeckman, B. (1992). The SWISS-PROT protein sequence data bank. *Nucl. Acids Res.* 20, 2019-2022

[18] Altschul, S. F., Boguski, M. S., Gish, W. G.& Wooton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genetics* **6**, 119-129.

[19] Pearson W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Science* 4, 1145-1160.

[20] Henikoff S. & Henikoff. J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915-10919.

[21] Wootton, J. C. & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* 17, 149-163.

[22] Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *JMB* **219**, 555-665.

[23] S. Karlin & S. F. Altschul. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *PNAS* **87**, 2264-2268.

[24] Bairoch, A. (1993) The PROSITE dictionary of sites and patterns in proteins, its current status. *Nucl. Acids Res.* **21**, 3097-3103.

[25] Nuoffer, C. & Balch, W. (1994). GTPase: multifunctional molecular switches regulating vesicular traffic. *Ann. Rev. Biochem.* 63, 949-990.

[26] Schuldiner, S., Shirvan, A. & Linial, M. (1995). Vesicular neurotransmitter transporters: from bacteria to man. *Physiological Reviews* **75**, 369-392.

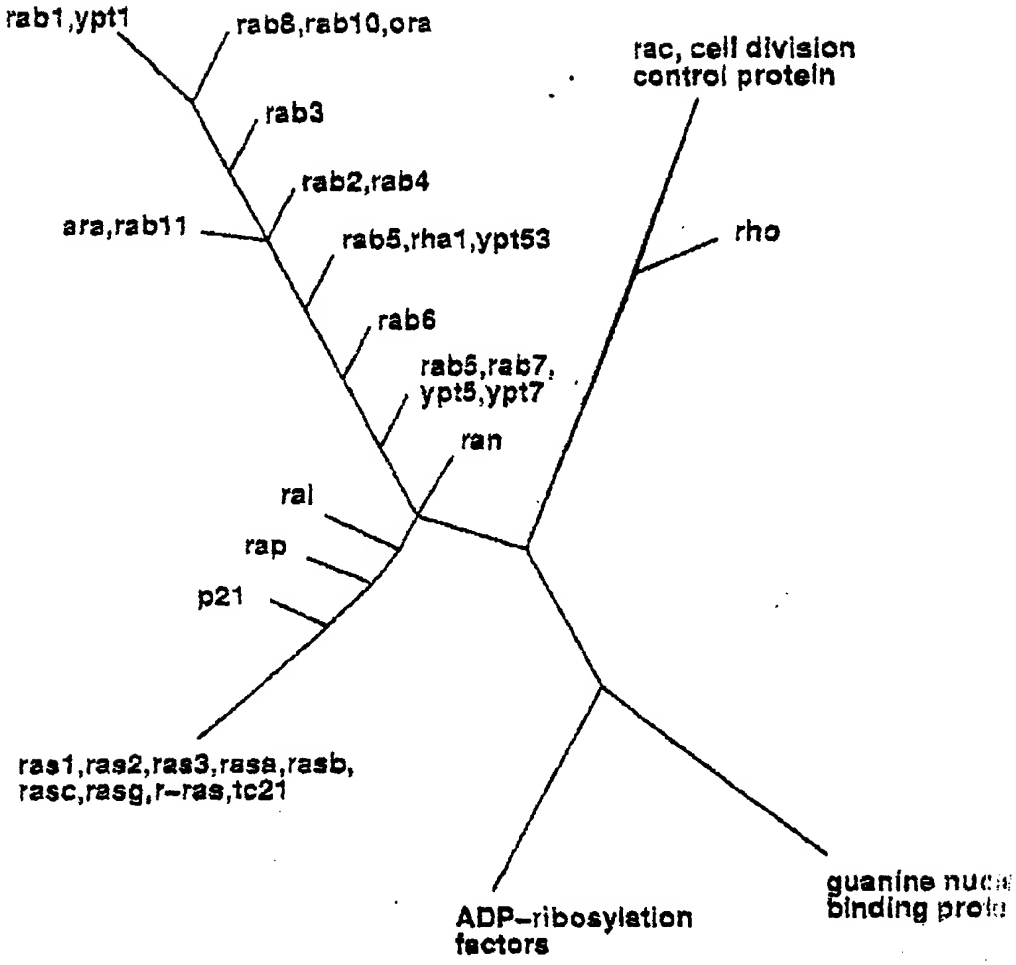



Figure 2: The small G-protein family. This family composed of few sub-families. The composition can be revealed based on the hierarchical organization we obtained. Total of 366 proteins were grouped together into isolated sets at different levels of confidence, to form a natural sub-classification within the family. At the level of $10^{-18}$ this family is linked with the ADP-ribosylation factors family and guanine nucleotide-binding proteins.

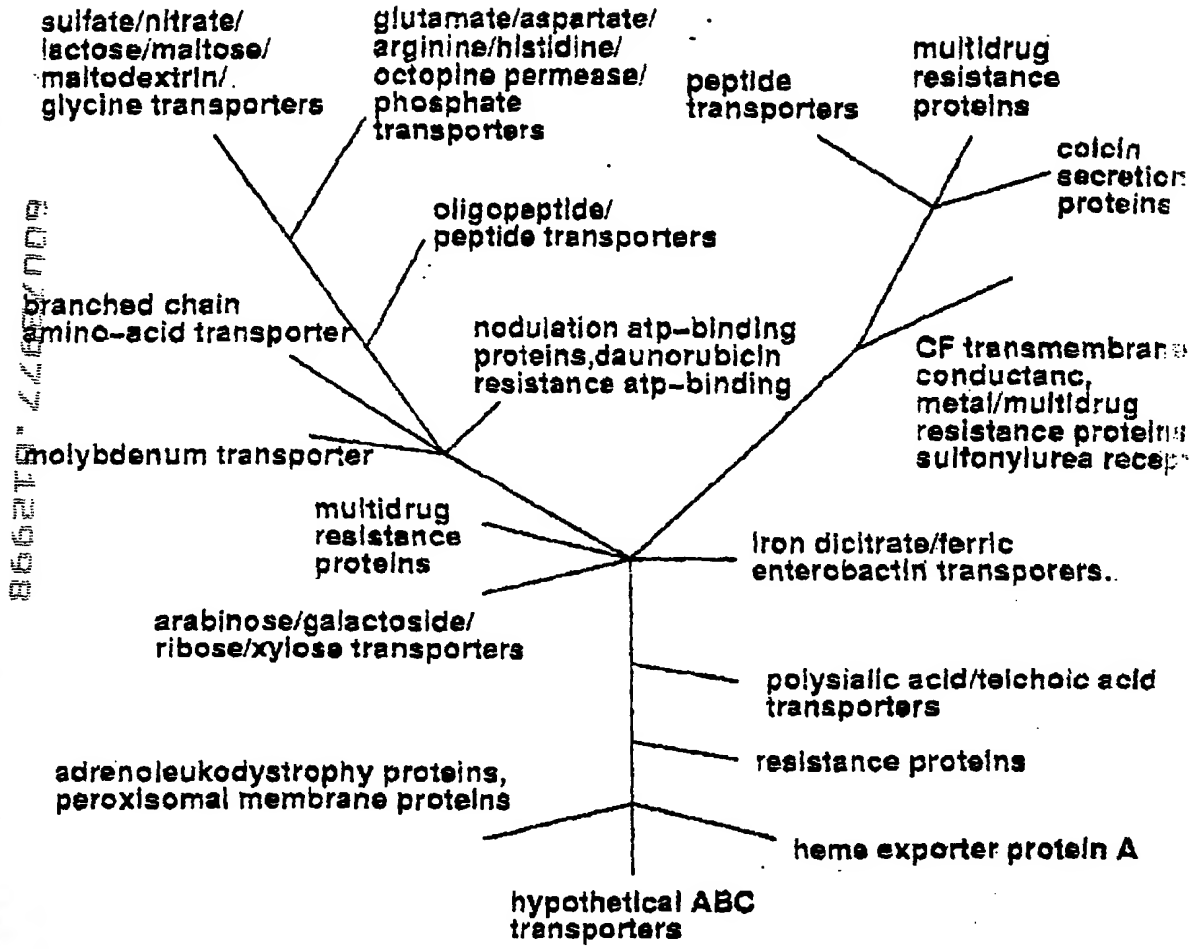



Figure 3: The ATP-binding transporters family.

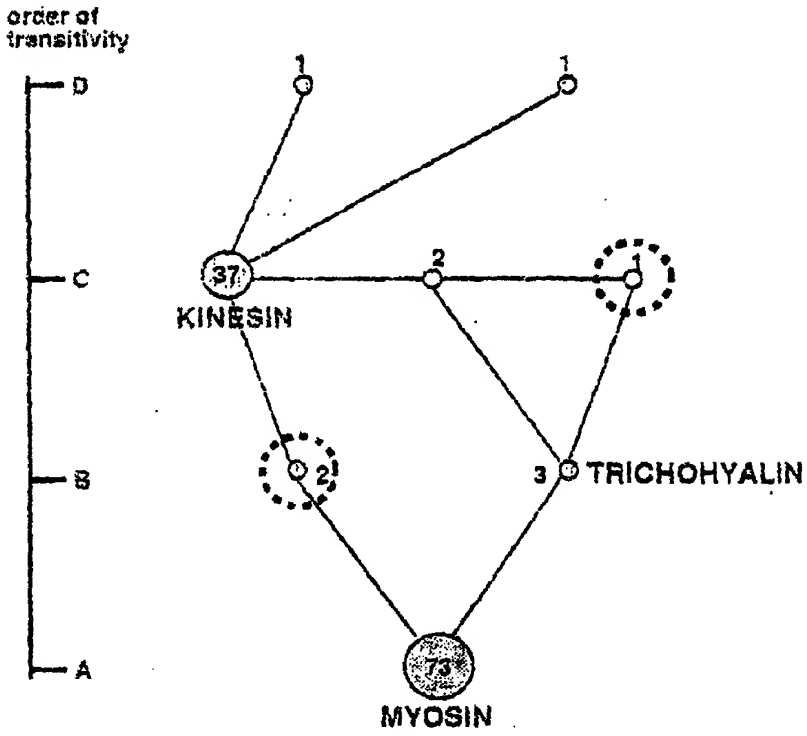



Figure 4: Super-family of motor proteins. Each circle stands for a connected component at threshold = $10^{-x}$. Circles' radii are proportionate to the component's size. The component's size appears next to the corresponding circle. The drawn edges appeared upon lowering the threshold to $10^{-30}$. The letters A-D indicate the order of transitivity. Each cluster is referred to by its order of transitivity A-D, and its position from left to right. Clusters which consists solely of hypothetical proteins are double circled.

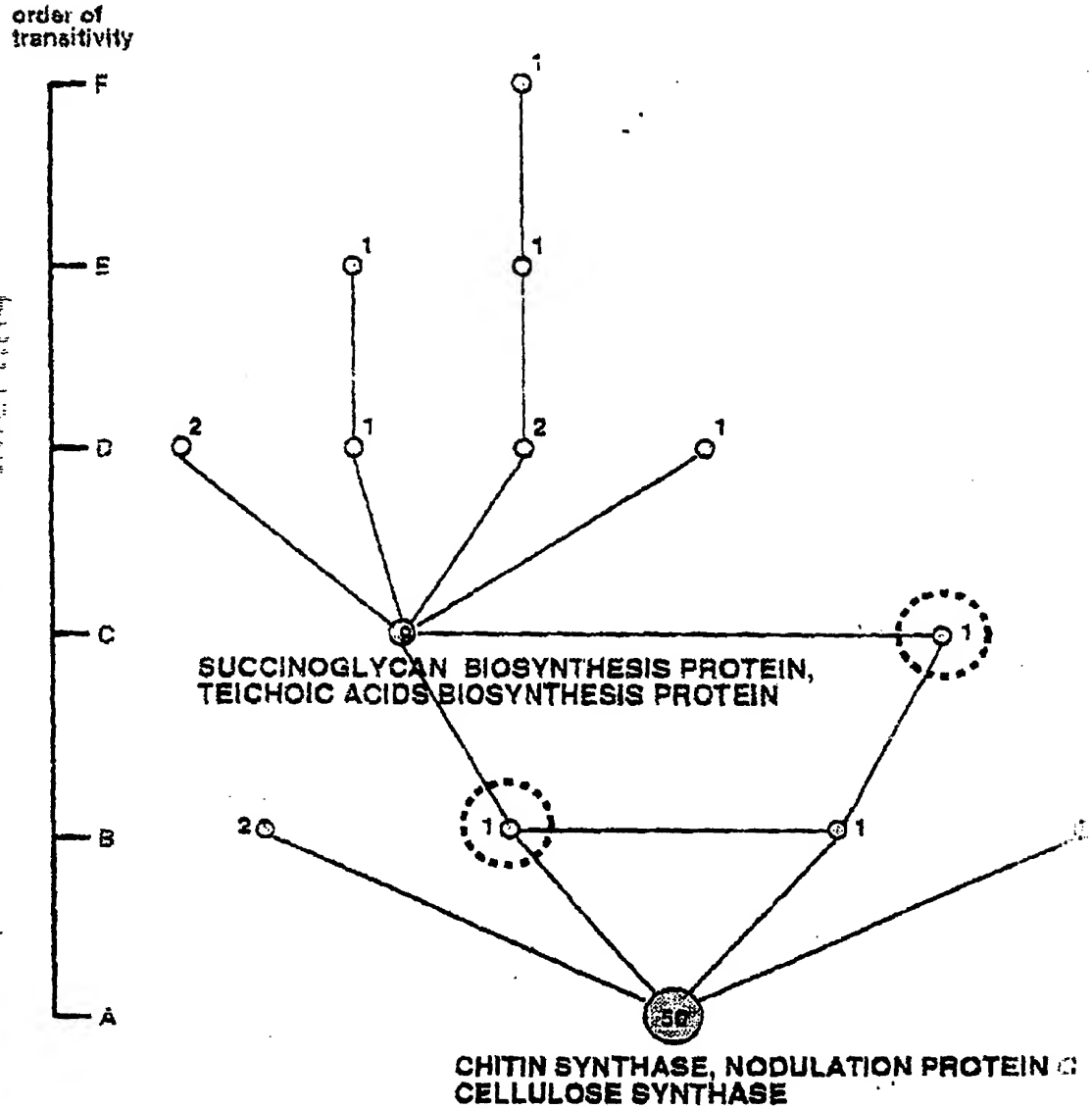
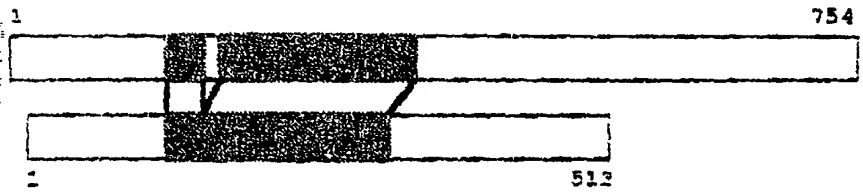



Figure 5: Proteins involved in the biosynthesis of complex sugars.




```
acsa_acexy   754 aa  <-> ychl_rhosc   513 aa     (score  76.0)

140  PLPDNVDDWPTVDIFIPTYDEQLSIVRLTVLGALGIDWPPDKVNVYILDDG..VRPEFEQFAKDCGALYIGRVDSSHAKK  213
     F P .     V : IP  D Q  : RL  .            . V : DDG  V  E  .     G : : R IS.   .
121  PRPMSIPSAADVTVVIPVCDNQAGVERLLPV.. .... LKDLTVIVVDDGSEVPVEPRRACPGIGTITVVRMESARGPS  191

213  GNLNHAIKRTSGDYILILDCDNIPTRAFLQIAMGWMVADRKIALMQTPHEFYSPDPFQRNLA..VGYRTPPEGNLFYGVI  291
       N  :     :: LD D IP  .L:: :G  .D  :AL.      . DP:  LA  : R.  :      :
193  AARNSGLRSAQTRFVAFLDSDVIPRAGWLELMLGHF.SDPGVALVAP..RIVALDPYGTALARYENMRSSLDLGRFEAAV  270

296  QDGNDFWDATFFCGSCAILRREAIESIGGFAVET.VTEDAHTALRMQRRGWSTAY  349
     . G.      :    . I.RR:      GF     V ED     H:Q  GW   Y
270  XSGSP...VAYVPSAAVIVRRDVALECNGFDLSLEVAEDVDFCWRLQAAGWRLRY  321

244 identity in 196 aa overlap

alignment parameters: blosum62 matrix, gap penalties -12,-2
```

Figure 6: Alignment of sw:acsa_acexy and sw:ychl_rhoso. The proteins are classified to sets A1 and B1 respectively in Fig. 5.

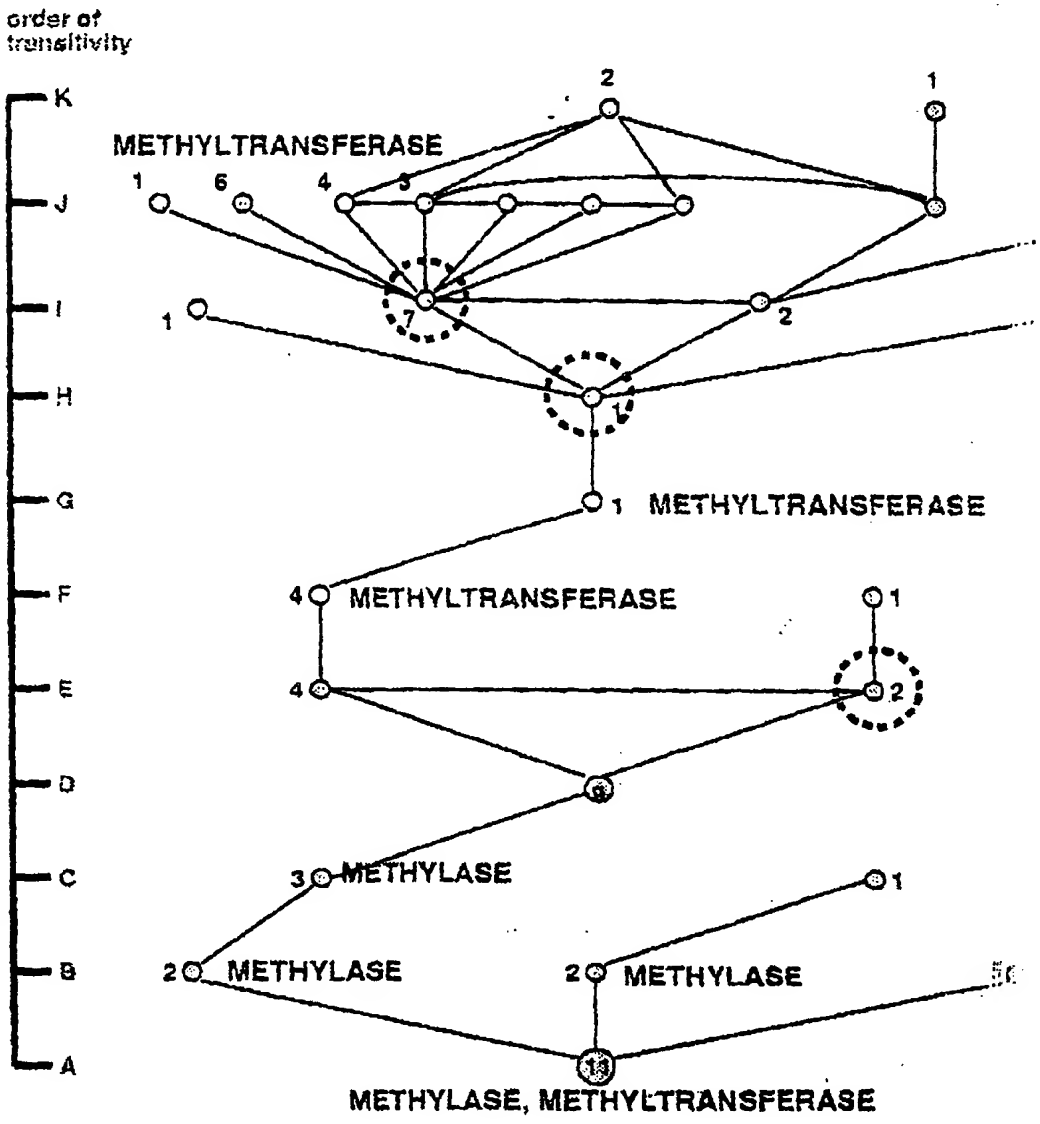
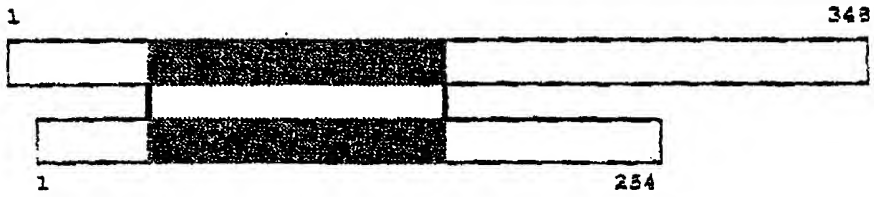



Figure 7: The methylases and methyltransferases. For details on the representation see Fig. 2.




```
hmt1_yeast   348 aa  <->  y912_haein   254 aa     (score  93.0)

58   KDKIVLDVGCGTGILSMFAAKHGAKHVIGVDMSS.IIEMA.KELVELNGFSDKITLLRGKLEDV.KLPFPKVDIIIS
     K K .LD.GCGTG        . GA  VIG D:S  .:E A K:L .   FS : .L     :E .  LP    DIL
44   KGKKLILDLGCGTGSMLQLYLERGAAKVIGTDLSEKMLEQAEKDLQKCGQFSGRFSLYHLPIEKLAELPEAHFDVII

135  GYFLLYESMMDTVLYARDHYLVEGGLIFPDKCSIHLAGLEDSQYK  179
       ::: E.     .    D     G LIF   .  I     E  .:.
124  FHYI..KNFPTLLSIIHDKLSSNGTLIFSQEMFITTCHKEGERME  168

30% identity in 120 aa overlap

alignment parameters: blosum62 matrix. gap penalties -12,-2
```

Figure 8: Alignment of sw:hmt1_yeast and sw:y912_haein. The proteins are classified to sets G1 and H1 respectively, in Fig. 7